A robust Bayesian meta-analytic approach to incorporate animal data into phase I oncology trials

Journal Title
XX(X):2–??

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Haiyan Zheng[1], Lisa V. Hampson[2], Simon Wandel[2]

Abstract

Before a first-in-man trial is conducted, preclinical studies are performed in animals to help characterise the safety profile of the new medicine. We propose a robust Bayesian hierarchical model to synthesise animal and human toxicity data, using scaling factors to translate doses administered to different animal species onto an equivalent human scale. After scaling doses, the parameters of dose-toxicity models intrinsic to different animal species can be interpreted on a common scale. A prior distribution is specified for each translation factor to capture uncertainty about differences between toxicity of the drug in animals and humans. Information from animals can then be leveraged to learn about the relationship between dose and risk of toxicity in a new phase I trial in humans. The model allows human dose-toxicity parameters to be exchangeable with the study-specific parameters of animal species studied so far or non-exchangeable with any of them. This leads to robust inferences, enabling the model to give greatest weight to the animal data with parameters most consistent with human parameters, or discount all animal data in the case of non-exchangeability of parameters. The proposed model is illustrated using a case study and simulations. Numerical results suggest that our proposal improves the precision of estimates of the toxicity rates when animal and human data are consistent, while it discounts animal data in cases of inconsistency.

Prepared using sagej.cls

Keywords
Bayesian hierarchical model; Historical data; Oncology; Phase I clinical trials; Robustness.

1 Introduction

There has been much recent interest in methods leveraging historical information for the design and interpretation of new clinical trials[1–5]. Information may be available from clinical trials, epidemiological studies, medical research or routine clinical practice. For example, patients randomised to standard of care or placebo in historical trials can be used to augment[6–8] or, in exceptional circumstances, substitute entirely[5] for the control arm of a new trial, thus enabling more ethical or smaller studies, or studies which learn more about the novel therapy. Methods for leveraging historical information have applications to trials in small or difficult to study populations, for example, paediatric trials[9] or studies of antibiotics for drug resistant pathogens[10]. In the context of early phase trials, Takeda and Morita[11] incorporate data from a completed phase I trial into a subsequent dose-escalation study performed in a different patient population. Cunanan and Koopmeiners[12] discussed possibilities of combining information across patient populations for a more accurate characterisation of the toxicity profile of a new compound in oncology.

When leveraging historical data, it is always possible that a conflict will emerge between the historical and the new trial data. In view of this, several approaches have been developed which downweight the historical data either to a degree that is fixed ahead of time or determined dynamically based on the extent of the observed prior-data conflict. Power priors[13] with a fixed exponent are examples of 'static priors'[1] while power priors with

[1]Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4SL, U.K.
[2]Novartis Pharma AG, Basel CH-4002, Switzerland

Corresponding author:
Lisa Hampson, Statistical Methodology and Consulting, Novartis Pharma AG, Basel CH-4002, Switzerland
Email: lisa.hampson@novartis.com

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

random exponents[14], commensurate priors[15,16], and meta-analytic analyses based on Bayesian hierarchical random-effects models[2,17,18] are examples of dynamic approaches.

This manuscript proposes a meta-analytic approach for leveraging animal data from preclinical studies in a phase I oncology trial which proceeds according to a Bayesian model-based design. So far numerous Bayesian procedures, based on one- or two-parameter models for the dose-toxicity relationship, have been proposed to use all accumulated data for informed decision making in phase I dose-escalation trials. Examples include the continual reassessment method[19,20], procedures implementing escalation with overdose control[21], and Bayesian decision theoretic approaches which make interim dose recommendations to maximise a gain function[22]. These designs have superior operating characteristics to the algorithmic 3+3 design[23], correctly identifying the true maximum tolerated dose (MTD) with higher probability and allocating a higher proportion of patients to this dose[24]. Whilst a one-parameter model may provide an adequate local approximation to the dose-toxicity relationship, when linking dose-toxicity relationships in animals and humans we will find it helpful to have a more complete description of how risk varies with dose, and so adopt a two-parameter Bayesian logistic regression model (BLRM)[25,26].

As far as we are aware, little has been written on quantitative methods for augmenting phase I clinical trials with animal data. Instead, attention has focused on using preclinical data to inform the choice of a safe starting dose for a phase I first-in-man trial[27,28]. A challenge one faces when synthesising data across species is that safe doses associated with an acceptable risk of toxicity in humans and different animal species may cover very different dosing intervals. To overcome this challenge, we will draw on techniques such as allometric scaling, to transform an animal dose onto an equivalent human scale, and robust meta-analytic combined analyses[18], so that we can rapidly discount information derived from preclinical studies in the event of a conflict between this and the observed human data.

The remainder of the paper is structured as follows. In Section 2, we propose a Bayesian meta-analytic model to borrow information from one or more animal species to humans. In

Section 3, we present a case study illustrating how the proposed hierarchical model can be used to analyse animal and human data at a single analysis. In Section 4, we use examples to explore how the model can be used to leverage animal data for interim decision making in a dose-escalation trial and describe the results of a simulation study evaluating trial operating characteristics in Section 5. Particular attention is given to evaluating the model's ability to react to a conflict between the animal data and accruing human data. We conclude in Section 6 with a discussion of possible extensions of the proposed methodology.

2 Incorporating preclinical animal data into a phase I first-in-man trial

Neuenschwander *et al.*[29] propose a Bayesian hierarchical model to augment a new phase I clinical trial with data from historical phase I studies, assuming that in each trial the relationship between dose and risk of toxicity follows a two-parameter logistic model. We describe below how this model can be extended to accommodate the case that existing data are observations from preclinical studies performed in one or more animal species.

Suppose that M preclinical studies have been performed in K animal species, with $K \leq M$, and let $\mathcal{S} = \{S_1, \ldots, S_K\}$ contain labels for the K species studied so far. Furthermore, we assume that a single animal species $\mathcal{A}_i \in \mathcal{S}$ was investigated in study i, for $i = 1, \ldots, M$. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ denote the vector listing the binary dose-limiting toxicity (DLT) outcomes (DLT or no DLT) of the n_i animals treated in study i. Finally, we suppose that the J_i doses contained in the set $\mathcal{D}_i = \{d_{i1}, \ldots, d_{iJ_i}; d_{it_1} < d_{it_2} \text{ for } 1 \leq t_1 < t_2 \leq J_i\}$ were evaluated in study i. Letting d_{ij} denote the jth dose for evaluation in study i and indexing dose by the subscript j, we suppose that r_{ij} out of the n_{ij} animals that received dose d_{ij} experienced a DLT, and model study i data as:

$$\begin{aligned} r_{ij}|p_{ij}, n_{ij} \sim \quad & \text{Binomial}(p_{ij}, n_{ij}), \quad \text{for } j = 1, \ldots, J_i \\ \text{logit}(p_{ij}) = \theta_{1i} + & \exp(\theta_{2i}) \log(\delta_{\mathcal{A}_i} d_{ij}/d_{\text{Ref}}) \end{aligned} \tag{1}$$

where p_{ij} denotes the DLT risk on dose d_{ij} and d_{Ref} is a reference dose invariant across studies defined below. We expect the dose-toxicity relationships of different species to

be more similar in terms of their slopes than locations[30], as shown in Supplementary Figure S1. With this in mind, for $k = 1, \ldots, K$, the term δ_{S_k} in Model (1) attempts to translate the doses administered to species S_k onto a common equivalent human dosing scale. Under this parameterisation, the intercept of the dose-toxicity model in study i is $\theta_{1i} + \exp(\theta_{2i}) \log(\delta_{\mathcal{A}_i})$, and therefore depends upon the animal species studied, whereas the slope is $\exp(\theta_{2i})$, and does not depend on the animal species. After translation, similar intervals of values should characterise acceptably safe doses in each animal species and humans. Therefore, θ_{1i} and θ_{2i} in (1) can be thought of, in an approximate sense, as the parameters that would have applied in study i had humans been studied rather than animal species $\mathcal{A}_i$. The translation factor δ_{S_k} reflects the relative potency of a compound in species S_k and humans; that is, if $\delta_{S_k} > 1$ ($0 < \delta_{S_k} < 1$), the same dose of a drug has a higher (lower) DLT risk in species S_k than in humans. A special case is $\delta_{S_k} = 1$, which implies a drug has a similar potency in species S_k and humans.

Allometric scaling[31,32] is a technique used to transform an animal dose into a human equivalent dose by adjusting for differences in size[33]. Specification of the translation factors in (1) can be informed by allometric scaling, assuming size-related differences in drug metabolism and pharmacokinetics explain differences in DLT risk between animals and humans given the same dose. In current practice, the translation factor is usually treated as a fixed constant. For example, to inform the selection of initial doses in human healthy volunteers, the FDA[27] advocates converting a no observed adverse event level in animals based on a body surface area correction factor with an allometric exponent of 0.67. However, there will usually be some uncertainty about the precise nature and extent of differences between humans and animals. To capture this uncertainty, we propose treating the translation factors, the δ_{S_k}'s, as random variables, which tends to reduce the amount borrowed from the animal data but also increases the robustness of our borrowing of information across species. We propose placing a log-normal prior on each δ_{S_k}. Table 1 lists log-normal priors specified using information from the FDA draft guideline *Estimating the Maximum Safe Starting Dose in Initial Clinical Trials for Therapeutics in Adult Healthy Volunteers*[27]; details on the derivation of these priors can be found in Appendix A.

Table 1. Log-normal priors for species-specific translation factors, $\delta_{\mathcal{A}_i} \sim LN(\lambda, \gamma^2)$, specified using body surface area (BSA) and body weight (BW) data documented in the FDA draft guidelines (FDA, 2005).

Species	BW (kg)		BSA (m^2)	HED in mg/kg		HED in mg/m^2	
	Reference	Working range		λ	γ	λ	γ
Mouse	0.02	(0.011, 0.034)	0.007	-2.562	0.298	1.050	0.283
Hamster	0.08	(0.047, 0.157)	0.016	-2.002	0.302	1.609	0.287
Rat	0.15	(0.080, 0.270)	0.025	-1.820	0.323	1.792	0.309
Ferret	0.30	(0.160, 0.540)	0.043	-1.669	0.323	1.943	0.309
Guinea pig	0.40	(0.208, 0.700)	0.050	-1.532	0.315	2.079	0.301
Rabbit	1.80	(0.900, 3.000)	0.150	-1.127	0.290	2.485	0.274
Dog	10	(5, 17)	0.500	-0.616	0.301	2.996	0.286
Primates:							
Monkeys	3	(1.400, 4.900)	0.250	-1.127	0.273	2.485	0.256
Marmoset	0.35	(0.140, 0.720)	0.060	-1.848	0.401	1.764	0.389
Squirrel monkey	0.60	(0.290, 0.970)	0.090	-1.715	0.269	1.897	0.252
Baboon	12	(7, 23)	0.600	-0.616	0.306	2.996	0.291
Micro-pig	20	(10, 33)	0.740	-0.315	0.284	3.297	0.268
Mini-pig	40	(25, 64)	1.140	-0.054	0.258	3.558	0.240

Model (1) assumes that for each k, translation factor δ_{S_k} applies across all studies performed in species S_k since δ_{S_k} is intended to capture intrinsic differences between species S_k and humans. We may consider refining this assumption if the different studies performed in species S_k focused on distinct subgroups, e.g., mature versus juvenile animals.

Now let $i^\star$ index the phase I first-in-man trial which will evaluate doses in the set $\mathcal{D}_{i^\star} = \{d_{i^\star 1}, \ldots, d_{i^\star J_{i^\star}}\}$. For completeness, we refer to humans as species $\mathcal{H}$ and define the label $\mathcal{A}_{i^\star} = \mathcal{H}$, denoting that humans will be studied in the new trial. Furthermore, let $\boldsymbol{\theta}_{i^\star} = (\theta_{1i^\star}, \theta_{2i^\star})$ denote the model parameters that will underpin the new trial. We model data from study $i^\star$ as:

$$\begin{aligned} r_{i^\star j} | p_{i^\star j}, n_{i^\star j} \sim \quad & \text{Binomial}(p_{i^\star j}, n_{i^\star j}), \quad \text{for } j = 1, \ldots, J_{i^\star} \\ \text{logit}(p_{i^\star j}) = \theta_{1i^\star} + & \exp(\theta_{2i^\star}) \log(d_{i^\star j}/d_{\text{Ref}}), \end{aligned} \tag{2}$$

where we stipulate $\delta_{\mathcal{A}_{i^\star}} = 1$ since human doses are already expressed on the common human dosing scale, and $d_{\text{Ref}} \in \mathcal{D}_{i^\star}$ is the same reference dose specified in (1).

Recall that if translation factors in (1) are appropriately specified, study-specific parameters will be on a common human dosing scale and there will be similarities between

the study-specific parameter vectors. We then stipulate

$$\boldsymbol{\theta}_i|\boldsymbol{\mu}_{\mathcal{A}_i}, \Psi \sim \text{BVN}(\boldsymbol{\mu}_{\mathcal{A}_i}, \Psi) \quad \text{with } \mathcal{A}_i \in \{S_1, \ldots, S_K\}, \tag{3}$$

and for each $S_k,\ k = 1, \ldots, K$,

$$\boldsymbol{\mu}_{S_k} = \begin{pmatrix} \mu_{1S_k} \\ \mu_{2S_k} \end{pmatrix} \quad \text{and} \quad \Psi = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}.$$

Variances in Ψ represent between-trial heterogeneity within an animal species. For increased borrowing of information between different animal species, we further assume the population means $\boldsymbol{\mu}_{S_1}, \ldots, \boldsymbol{\mu}_{S_K}$ are exchangeable. A bivariate normal 'supra-species' random-effects distribution is stipulated as follows. For each $S_k,\ k = 1, \ldots, K$,

$$\boldsymbol{\mu}_{S_k}|\boldsymbol{m}, \Sigma \sim \text{BVN}(\boldsymbol{m}, \Sigma), \tag{4}$$

with

$$\boldsymbol{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \kappa\sigma_1\sigma_2 \\ \kappa\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

The random-effects distribution in (4) accounts for between-species differences in average dose-toxicity model parameters. We note such differences may arise due to misspecification of one or more translation factors δ_{S_k}; for example, if there are size-dependent and size-independent differences between an animal species and humans, the latter may not be completely captured by δ_{S_k}, but can be addressed by variances in Σ.

The Bayesian hierarchical model for the preclinical data is completed by specifying prior distributions for the hyperparameters, where we implement the model setting

$$\begin{aligned} m_1 &\sim N(v_1, s_1^2), \quad m_2 \sim N(v_2, s_2^2), \\ \tau_1 &\sim HN(z_1), \quad \tau_2 \sim HN(z_2), \quad \rho \sim U(-1, 1), \\ \sigma_1 &\sim HN(c_1), \quad \sigma_2 \sim HN(c_2), \quad \kappa \sim U(-1, 1). \end{aligned} \tag{5}$$

Here, $HN(z)$ denotes a half-normal distribution formed by truncating a $N(0, z^2)$ distribution to cover the interval $(0, \infty)$. Although it will not be considered here, one could allow the between-study variances in Ψ to vary across species.

We have yet to say how we relate the human study-specific parameter vector $\boldsymbol{\theta}_{i^\star}$ to the animal study-specific parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$. We require robust borrowing of information across species, meaning that we should downweight information from species with dose-toxicity model parameters dissimilar to those in humans, and discount all preclinical data if no animal species appears similar to humans. Then, for each $k = 1, \ldots, K$, we stipulate

$$\boldsymbol{\theta}_{i^\star} | \boldsymbol{\mu}_{S_k}, \Psi \sim \text{BVN}(\boldsymbol{\mu}_{S_k}, \Psi) \quad \text{with prior probability } w_{S_k},$$

so that w_{S_k} represents the prior plausibility that $\boldsymbol{\theta}_{i^\star}$ is exchangeable with the study-specific parameters in species S_k. Note that we define exchangeability at the level of the study-specific model parameters since $\boldsymbol{\theta}_{i^\star}$ is a study-specific, rather than population mean, parameter. For robust inferences about $\boldsymbol{\theta}_{i^\star}$, we stipulate

$$\boldsymbol{\theta}_{i^\star} \sim \text{BVN}(\boldsymbol{m}_0, R_0) \quad \text{with prior probability } w_R,$$

where $w_R = 1 - \sum_{k=1}^{K} w_{S_k}$ is a prior non-exchangeability weight and $\text{BVN}(\boldsymbol{m}_0, R_0)$ is a weakly informative prior distribution. In practice, specification of $w_{S_1}, \ldots, w_{S_K}$ will require the input of subject-matter experts such as translational scientists or pharmacologists. The robust hierarchical model is fitted using Markov chain Monte Carlo, and thus can be implemented with software such as OpenBUGS[34].

We note that adding a 'supra-species' level to the Bayesian hierarchical model in equation (4) allows for increased, but robust, borrowing of information across species. When all the $\boldsymbol{\theta}_i$s are similar to both each other and $\boldsymbol{\theta}_{i^\star}$, we can borrow strength across the related animal species to estimate the animal population mean parameters with greater precision, and thus gain additional precision for estimating $\boldsymbol{\theta}_{i^\star}$. Such borrowing is robust in the sense that if we place weakly informative priors on elements of Σ and find that, say, study-specific

Table 2. Ocular toxicities and general DLTs due to all cause observed during a phase I first-in-man trial of AUY922. Estimated risks are naive maximum likelihood estimates based on the pooled human data alone.

	Dose (mg/m^2)								
	$d_{i^\star 1}$	$d_{i^\star 2}$	$d_{i^\star 3}$	$d_{i^\star 4}$	$d_{i^\star 5}$	$d_{i^\star 6}$	$d_{i^\star 7}$	$d_{i^\star 8}$	$d_{i^\star 9}$
	2	4	8	16	22	28	40	54	70
Number of patients	3	3	4	6	11	8	16	18	24
Number of ocular AEs	0	0	0	0	0	0	0	0	2
Ocular AE risk	0.001	0.002	0.004	0.008	0.012	0.015	0.023	0.033	0.045

parameters of only one animal species are similar to $\boldsymbol{\theta}_{i^\star}$, posterior distributions for elements of Σ will place larger probability mass on large between-species variances. This leads to less borrowing across animal species to estimate the $\boldsymbol{\mu}_{S_k}$s, and we tend to borrow from the most relevant animal species to learn about $\boldsymbol{\theta}_{i^\star}$.

3 Illustrative example

In this section, we apply the proposed Bayesian hierarchical model to a retrospective example, synthesising preclinical and clinical ocular toxicity data on AUY922, an experimental compound intended to treat cancer[35,36].

3.1 Animal data

The safety profile of AUY922 was evaluated in several preclinical studies prior to its evaluation in humans. For this compound, ocular adverse events (AEs) were thought to potentially occur in humans. Therefore, the risk of this type of event was investigated in four studies performed in a total of 152 Wistar and Brown Norway rats[35], which we will hereafter refer to as 'rats'. The ocular AE data are displayed in Figure 1. The first two datasets are outcomes from Studies 1 and 2 reported in Roman *et al.*[35]. Since Study 1 involved male and female rats but Study 2 involved only males, we use only the male rat data from Study 1. It was not possible to extract the ocular AE data of Studies 3 and 4 from the same preclinical paper[35]. Therefore, Figure 1 shows simulated, but plausible, data for these studies instead (slight modifications to the doses for these studies have also been made so that we will have data on various doses to fit the logistic model for rats). Data from the phase I study of AUY922 were published in Sessa *et al.*[36] and are listed in Table 2. During the phase I trial, doses from the set $\mathcal{D}_{i^\star} = \{2, 4, 8, 16, 22, 28, 40, 54, 70\}$ mg/m^2

were available for administration. The dose-escalation study was performed according to a BLRM-guided procedure monitoring DLTs, defined as the occurrence of any clinically relevant drug-related AE or abnormal lab value. Ocular AEs were also reported separately in the phase I clinical trial paper[36].

[Insert Figure 1]

In Section 3.2, we describe what would have been the prior predictive distributions for the risk of an ocular AE in the phase I trial given the rat data. In this example, since animal data were available from one species, we implement the robust Bayesian hierarchical model from Section 2 setting $K = 1$. We note that our model can accommodate the special case that $K = 1$ if weakly informative priors are adopted for diagonal elements of Σ. In Section 3.3, we refit the hierarchical model to incorporate both the rat and human data collected during the AUY922 phase I trial, and derive posterior distributions for the risk of an ocular AE in the human trial.

3.2 Prior predictive distributions for the risk of ocular toxicity in a phase I trial

We use the four rat datasets to fit the hierarchical model proposed in Section 2, setting $d_{\text{Ref}} = 28$ mg/m^2 and using the following priors. We set $m_1 \sim N(-1.099, 1.98^2)$ which implies a 95% prior credible interval for the risk of toxicity at 28 mg/m^2 is 0.007 to 0.942 and prior median 0.250. Furthermore, we set $m_2 \sim N(0, 0.99^2)$ to accommodate flat to very steep dose-toxicity curves. These are weakly informative priors that place probability mass on plausible values of the model parameters[37]. A similar approach is used to specify the parameters of the BVN($\boldsymbol{m}_0, R_0$) non-exchangeability prior. For the variance parameters, we set $\tau_1 \sim HN(0.5)$ assuming substantial variability between the study-specific θ_{i1}s, and $\tau_2 \sim HN(0.25)$, assuming a smaller degree of variability between the slopes of study-specific dose-toxicity curves. Larger values are specified for the half-normal priors placed on σ_1 and σ_2 to preclude giving definitive information. More details are given in Appendix B on the prior specification of hyperparameters. Finally, we

Table 3. Summaries of marginal predictive priors derived from the rat data setting $w_R = 0.5$. Also reported are the parameters of the Beta(a, b) approximates used for ESS calculations.

	Dose (mg/m^2)									
	$d_{i^\star 1}$	$d_{i^\star 2}$	$d_{i^\star 3}$	$d_{i^\star 4}$	$d_{i^\star 5}$	$d_{i^\star 6}$	$d_{i^\star 7}$	$d_{i^\star 8}$	$d_{i^\star 9}$	$d_{i^\star 10}$
	2	4	8	16	22	28	40	54	70	140
Prior means	0.062	0.080	0.107	0.150	0.179	0.209	0.259	0.300	0.335	0.424
Prior std dev.	0.148	0.166	0.189	0.219	0.237	0.254	0.284	0.305	0.317	0.330
ESS	1.5	1.5	1.6	1.6	1.6	1.5	1.3	1.2	1.2	1.2
a	0.1	0.1	0.2	0.2	0.3	0.3	0.4	0.4	0.4	0.5
b	1.6	1.5	1.5	1.4	1.3	1.2	1.0	0.9	0.8	0.7

stipulate $\delta_{\text{Rat}} \sim LN(1.792, 0.309^2)$.

[Insert Figure 2]

Figure 2A summarises predictive priors of the risk of an ocular AE in humans in the new phase I trial. Priors are derived at each human dose under a range of non-exchangeability weights. Each predictive prior is summarised by its median and 95% credible interval. Setting $w_R = 0$, prior predictive distributions are derived assuming full exchangeability between human and animal study-specific parameters. Increasing w_R to 0.5 suggests a large degree of prior skepticism about the plausibility of exchangeability. Setting $w_R = 1$ means we discard the rat data entirely so that the prior for $\boldsymbol{\theta}_{i^\star}$ is the weakly informative operational prior. Figure 2B further summarises priors derived setting $w_R = 0.5$ by three interval probabilities. We characterise the predictive prior for each dose by the probability: (i) of underdosing, which is said to occur if the DLT risk is less than 0.16; (ii) that the DLT risk lies in the target interval [0.16, 0.33); and (iii) of overdosing, which is said to occur if the DLT risk lies in the interval [0.33, 1][26]. Figure 2C presents the predictive prior probability densities of DLT risks on two low doses, 4 and 8 mg/m^2, when $w_R = 0.5$. Such visualisations may be useful for teams to consider when selecting the starting dose for a phase I trial.

To calculate the effective sample size (ESS)[38] of the predictive prior for the risk of an ocular AE on each human dose in the phase I trial, we approximate each prior by a Beta(a, b) distribution with parameters chosen to match the first two moments of the prior. The

ESS is then found as $(a + b)$. This follows because a Beta(a, b) prior can be thought of as representing opinion on the risk of an ocular AE after a out of $(a + b)$ patients allocated to a dose experience a toxicity, assuming nothing was known about the risk *a priori*[39]. After approximation, ESSs of predictive priors derived under $w_R = 0.5$ are listed in Table 3. The information represented by each prior is equivalent to that would be obtained from approximately 1.2 – 1.6 human patients, and so it is clear that there is heavy discounting of the preclinical data from 152 rats.

3.3 Synthesis of rat and human data on the termination of the phase I trial

We now apply the proposed methodology to synthesise ocular AE data from both rats and the data from humans available on termination of the phase I trial. Posterior distributions for the risk of an ocular AE on each human dose derived under models with different non-exchangeability weights are summarised in Figure 2D. Figures 2E-F summarise the posteriors derived setting $w_R = 0.5$.

With $1 - w_R = 0.5$, the posterior probability of exchangeability between rat and human study-specific parameters increases from the prior value of 0.5 to 0.82, suggesting that rat and human ocular AE data are more consistent than expected. Posterior median probabilities of an ocular AE in the human phase I study at doses 70 and 140 mg/m^2 are 0.048 (95% CI: [0.014, 0.118]) and 0.096 (95% CI: [0.025, 0.329]), respectively. These are slightly more cautious and narrower than the posterior medians and 95% CIs that would have been obtained had we discarded the rat data entirely from our inferences. Setting $w_R = 1$, the posterior median probabilities of an ocular AE (95% CIs) at 70 mg/m^2 and 140 mg/m^2 are 0.045 [0.010, 0.137] and 0.087 [0.015, 0.558], respectively. The marginal posterior distributions of the risk of an ocular AE in the human trial at the two highest doses when $w_R = 0.5$ are shown in Figure 2F.

4 Leveraging animal data in a phase I dose-escalation trial

In this section, we illustrate how our Bayesian hierarchical model can be used to leverage animal data for decision making in a hypothetical phase I dose-escalation trial.

4.1 Trial design and determination of a safe starting dose

Suppose that a phase I dose-escalation study, labelled $i^\star$, is to be performed to estimate the MTD in humans, defined here as the dose associated with a risk of a DLT (of any type) of 25%. During the phase I trial, doses (in mg/m^2) from the set $\mathcal{D}_{i^\star} = \{2, 4, 8, 16, 22, 28, 40, 54, 70\}$ will be available for administration. We suppose that at the time of designing the dose-escalation study, three studies have been conducted in dogs. Simulated data from these hypothetical studies are presented in Figure S2 in the Web-based Supplementary Materials. In our notation, these data are represented by $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3$. We analyse these data by fitting the Bayesian hierarchical model with priors setting $\tau_1 \sim HN(0.25)$ and $\tau_2 \sim HN(0.125)$, to assume moderate to small between-study variabilities for θ_{1i} and θ_{2i}, respectively, and $\delta_{\text{Dog}} \sim LN(2.996, 0.286^2)$. Priors for other parameters remain unchanged from Section 3.2.

[Insert Figure 3]

Figure 3A summarises the prior predictive distributions for the DLT risk in the new human study $i^\star$ on each dose in $\mathcal{D}_{i^\star}$. Setting $w_R = 0.3$, the median of the predictive prior for the DLT risk on dose 22 mg/m^2 is 0.252, with 95% CI [0.011, 0.800]. Figure 3B summarises these prior predictive distributions by presenting probabilities that the DLT risk lies in each of the three intervals (underdosing; target; and overdosing) defined in Section 3.2. We see that doses up to and including 16 mg/m^2 are associated with a prior predictive probability of overdosing of less than 25%. All hypothetical phase I dose-escalation studies start by allocating the first cohort 4 mg/m^2, with the possibility to de-escalate to 2 mg/m^2. On the basis of the dog data and our prior beliefs about their relevance with human data, 4 mg/m^2 appears very safe with $\mathbb{P}(p_{i^\star 2} < 0.1 \mid \boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3) = 0.790$.

4.2 Hypothetical dose-escalation studies

Suppose that patients enter the phase I trial in cohorts of size three and that all patients within a cohort receive the same dose. After each cohort has been treated and observed, an interim analysis is performed, at which point all dog and human data are analysed to recommend a

dose for the next cohort. Cohort $h = 1$ receives 4 mg/m^2. Letting $Y_{i^\star}^{(h-1)}$ denote the vector of outcomes from the first $(h-1)$ human cohorts, the escalation rule recommends that cohort $h = 2, \ldots$ receives dose

$$d_{\text{sel}}^{(h)} = \max\{d_{i^\star j} \in \mathcal{D}_{i^\star} : \mathbb{P}(p_{i^\star j} \geq 0.33 | \boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3, \boldsymbol{Y}_{i^\star}^{(h-1)}) \leq 0.25\}. \tag{6}$$

Dose recommendations are also subject to the additional constraint that escalation is restricted to a maximum two-fold increase in the current dose. For the dosing set considered here, this constraint implies that if the previous cohort received a dose $d_{i^\star j} \leq 16$ mg/m^2, the next cohort can escalate by at most one dose level so long as the overdose control criterion is satisfied.

[Insert Figure 4]

Figure 4 summarises the progress of eight hypothetical phase I trials run with simulated data, which are analysed using the proposed hierarchical model setting $w_R = 0.3$. Figure 4A traces dose-escalation recommendations while Figure 4B records how the posterior probability of exchangeability between the new human and dog study-specific parameters evolves as the study progresses. For reasons of parsimony, we monitor each simulated trial until any dose is recommended for a third time.

In examples 1 to 5, data were simulated so as to be largely consistent with the prior opinion illustrated in Figure 4A (when $w_R = 0.3$) that the DLT risk in humans given 22 mg/m^2 in the new trial will be close to 25%, while we are confident that the risks of toxicity on 2, 4 and 8 mg/m^2 will all be well below 33%. This consistency leads to higher posterior exchangeability probabilities, as shown in Figure 4B. In contrast, examples 6 to 8 represent cases where there is a conflict between the human data and what was anticipated based on the analysis of the dog data.

In examples 6 and 7, the simulated human data appear consistent with a higher DLT risk at lower doses than what was predicted *a priori*. In example 6, one out of three patients in the second cohort treated with 8 mg/m^2 experienced a DLT; we escalated to administer 16 mg/m^2 to the third cohort and all three patients experienced a DLT. Preclinical data from dog studies were then discounted, with a drop in the posterior probability of exchangeability from 0.810 to 0.358. A similar response to early observations of DLTs on low doses was seen in example 7.

In example 8, the first DLT was observed only after dosing reached 54 mg/m^2, so that the DLT risk at high doses appeared to be lower than what was predicted on the basis of the dog data. This prior-data conflict resulted in the posterior probability of exchangeability shifting from its prior value of 0.7 to 0.266 once data were available from the first six cohorts. Since the predictive prior derived from the dog data suggested that the human MTD in the new study would likely lie in the neighbourhood of 22 mg/m^2, it is not surprising that dose escalation slowed down as we approached this dosing range. After completion of the forth cohort, posterior probabilities of overdose at doses 28 and 40 mg/m^2 were 0.085 and 0.293, respectively. Thus, despite the fact that no human DLTs had been observed, the procedure repeated administration of 28 mg/m^2 to the fifth cohort.

5 Simulation study

We performed a simulation study to evaluate the operating characteristics of a phase I dose-escalation procedure. We simulate trials which proceed sequentially, recruiting patients in cohorts of size three. Trials proceed using the Bayesian hierarchical model of Section 2 to leverage the dog data illustrated in Figure S1. The preclinical data are held fixed in the analysis of all simulated trials. At each analysis, we fit the Bayesian hierarchical model with four choices for w_R:

- Model A: Full exchangeability between the $\boldsymbol{\theta}_i$s and $\boldsymbol{\theta}_{i^\star}$ ($w_R = 0$);
- Model B: High level of prior confidence in the exchangeability assumption ($w_R = 0.3$);
- Model C: Prior ambivalence about the exchangeability assumption ($w_R = 0.5$);
- Model D: No borrowing of information from the dog data ($w_R = 1$).

Table 4. Simulation scenarios for the true probability of DLT in humans and MAP priors with median and 95% credible intervals derived from the dog data setting $w_R = 0$. The figure in bold indicates the target dose closest to the true MTD.

	Dose (mg/m^2)								
	$d_{i^\star 1}$ 2	$d_{i^\star 2}$ 4	$d_{i^\star 3}$ 8	$d_{i^\star 4}$ 16	$d_{i^\star 5}$ 22	$d_{i^\star 6}$ 28	$d_{i^\star 7}$ 40	$d_{i^\star 8}$ 54	$d_{i^\star 9}$ 70
Probability of DLT in humans									
Scenario 1	0.08	0.16	**0.25**	0.35	0.41	0.45	0.52	0.58	0.63
Scenario 2	0.01	0.04	0.11	**0.25**	0.35	0.44	0.55	0.65	0.73
Scenario 3	0.03	0.05	0.10	0.16	**0.25**	0.32	0.40	0.48	0.55
Scenario 4	0.001	0.005	0.03	0.10	0.16	**0.25**	0.38	0.50	0.60
Scenario 5	0.01	0.02	0.05	0.08	0.11	0.14	**0.25**	0.37	0.47
Scenario 6	0.003	0.006	0.01	0.02	0.05	0.08	0.15	**0.25**	0.37
Scenario 7	**0.25**	0.42	0.60	0.75	0.82	0.88	0.91	0.94	0.97
Scenario 8	0.001	0.005	0.01	0.02	0.04	0.05	0.10	0.16	**0.25**
Prior medians and 95% credible intervals									
	0.02 (0.00, 0.14)	0.04 (0.00, 0.20)	0.10 (0.02, 0.29)	0.19 (0.05, 0.43)	0.26 (0.09, 0.51)	0.32 (0.12, 0.57)	0.42 (0.19, 0.68)	0.51 (0.26, 0.77)	0.59 (0.33, 0.83)

Interim dose recommendations are made according to rule (6), with the same caveats as described in Section 4.2. Trials end: i) once 45 patients have been treated and observed; or ii) at any interim analysis if the lowest dose is found to be excessively toxic, that is, the trial stops at interim analysis $(h-1)$ if $\mathbb{P}(p_{i^\star 1} \geq 0.33 \mid \boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3, \boldsymbol{Y}_{i^\star}^{(h-1)}) > 0.25$. These two subsets of simulated trials will later be referred to as *completed* and *stopped early* trials, respectively.

We consider eight different simulation scenarios, shown in Table 4, for the true dose-toxicity relationship in the new phase I trial. These include scenarios which are consistent with the predictive prior derived from the dog data, as well as scenarios in which the drug is more (or less) toxic than would be expected from the dog data. For each scenario and model, results are based on 2000 simulated trials.

Define $\tilde{p}_{i^\star j}$ as the point estimate (posterior median) of the DLT risk on dose $d_{i^\star j} \in \mathcal{D}_{i^\star}$. Then at the end of a *completed* trial, we estimate the MTD as:

$$\hat{d}_{\mathrm{M}} = \arg \underset{d_{i^\star j} \in D'_{i^\star}}{min} |\tilde{p}_{i^\star j} - 0.25|,$$

where $\mathcal{D}'_{i\star} \subseteq \mathcal{D}_{i\star}$ comprises all the doses that have been administered to humans during the trial and satisfy the probabilistic overdose criterion. In each simulation scenario, we record the percentage of studies which identify each dose as the MTD. We also record the percentage of trials which *stop early* without a MTD declaration. Furthermore, averaging across the 2000 simulated trials, we report the average number of patients allocated to each dose.

[Insert Figure 5]

Figure 5 compares dose-escalation procedures implemented using Models A – D in terms of the percentage of trials which correctly select the MTD (PCS), the percentage of trials which stop early for safety; and the average number of patients allocated to the true MTD. Procedures underpinned by Models B and C perform reasonably well across all eight simulation scenarios. In cases where there is a strong prior-data conflict, for example in Scenarios 7 and 8, procedures based on Model C tend to slightly outperform those based on Model B. When there is prior-data consistency, such as in Scenario 3, the relative performances are reversed, although differences between the models remain small across all scenarios.

Comparing Models B and C with Model D, we see that by leveraging the dog data we can make gains for the PCS and average number of patients assigned to the true human MTD when the dog data are predictive of DLT risks in the new phase I trial. For example, we see an increase in PCS of at least 12.9% in Scenario 3. However, Model D clearly outperforms Models B – C in Scenario 8, in terms of the average number of patients allocated to the true MTD, although smaller differences emerge in terms of the PCS.

Comparing Models B and C with Model A, we see the advantages of robustification in Scenarios 6 and 8, where the assumption of full exchangeability leads to underestimation of the MTD, and allocation of a higher average number of patients to lower doses. The impact of robustification when an assumption of exhangeability is appropriate is seen in Scenario

3, when PCS decreases from 55.6% ($w_R = 0$) to 45.8% ($w_R = 0.5$).

[Insert Figure 6]

For analysis Models A-C, we estimate δ_{Dog} by the median of its posterior distribution at the end of each *completed* trial. Figure 6 compares in each simulation scenario the distribution of posterior median estimates of δ_{Dog} with the prior median represented by the solid horizontal line. The deviation of the posterior median estimate from the prior median reflects the prior-data conflict. For example, in Scenarios 1 and 2 when preclinical data under-predict the potency of the drug in the phase I study, the posterior estimates of δ_{Dog} tend to decrease from the prior estimate to adjust for this emerging conflict. Treating δ_{Dog} as a random variable provides a mechanism to respond to prior-data conflicts and therefore leads to more robust borrowing of information across species. The posterior estimates of δ_{Dog} in Scenario 7 appear to be less dispersed, because few trials were completed in this highly toxic scenario. Within a scenario, the size of the shift in posterior estimates decreases across Models A – C. As w_R increases, the need to respond to the prior-data conflict by updating δ_{Dog} becomes less as the prior weight on the exchangeability scenario decreases.

Another interesting evaluation is to compare two variants on Models A – C treating δ_{Dog} as either a random variable or a fixed constant adopted in current practice. The optimal non-parametric benchmark design[40] is also considered for comparison to assess potential gains of leveraging preclinical data in different simulation scenarios. Given different analysis models, we also investigated the bias, mean squared error and coverage probability of the central 95% credible interval of the posterior estimate of the DLT risk at the true MTD. Results of these assessments are available in Figures S3 and S4 in the Web-based Supplementary Materials. Furthermore, we have re-run selected simulations setting $\tau_2 \sim HN(0.25)$ instead of $\tau_2 \sim HN(0.125)$. As expected, a larger value of the scale parameter leads to reduced borrowing of information from the preclinical data while general conclusions for the comparison of different models are unchanged. Finally, we notice in practice there are situations where a phase I trial may be implemented with early stopping

rules to declare the MTD. We thus consider dose-escalation procedures based on Models A – D with rules permitting early stopping when specified conditions are met. Operating characteristics are summarised in Figure S5 in the Supplementary Materials.

6 Discussion

Bayesian meta-analytic approaches provide a framework to augment a clinical trial with historical data. In this paper, we have proposed a robust Bayesian hierarchical model to augment a first-in-man trial with data from preclinical toxicology studies in animals. The novelty of this approach is two-fold: First, we translate the dose-toxicity curves from different animal species onto the human scale, which allows us to adequately combine the information from animals and humans. Second, the translation factor used for scaling is a parameter with uncertainty in the model, and we estimate this parameter more and more precisely as data in humans (or from different species) accumulate. In our opinion, both of these points are important to leverage the dose-toxicity information from preclinical to clinical studies in a transparent and statistically efficient way. The simulations presented in Section 5 show that the proposed methodology enables robust borrowing of information from animals to humans, and is responsive to prior-data conflicts. We note, however, that when there is a substantial prior-data conflict, using our approach may lead to a decrease in precision of the estimate, regardless of how small the prior weight assigned to the animal data is.

In practice, often only few animal safety studies are conducted prior to the phase I clinical trial. Our data examples and simulation study presented in Sections 3 – 5 have preclinical data collected from only one animal species, presenting applications of our methodology in quite restrictive cases with only limited preclinical information. Additional simulations have been performed (results not reported here) to verify the performance of the meta-analytic model for cases that $K = 2$ and $K = 3$. These supported similar conclusions to those shown in this paper, namely that borrowing of information from animals to humans is robust and is led by data from the most relevant animal species. Having a larger number of preclinical studies involving multiple animal species is potentially advantageous for

estimating the variance parameters that are associated with between-study and between-species heterogeneity. We would also like to add a note on the exchangeability of the population mean parameters $\boldsymbol{\mu}_{S_k}$ when $K > 1$: learning about the variance parameters in the 'supra-species' level would be important to facilitate sharing of information between different species to an appropriate extent. Another critical concern may be the number of doses tested in the animals studies. While making inferences is always possible in a Bayesian model regardless of the amount of data, for a meaningful use of our approach we recommend that toxicity data need to be generated on at least two different doses in minimally one animal (or human) study before the new phase I clinical trial is conducted.

High quality preclinical data are essential to design an ethical phase I clinical trial[41,42]. Current approaches to using animal data culminate in a safe starting dose for a phase I clinical trial. This underutilises the toxicity data accumulated from the animal studies. To our knowledge, this paper represents a first proposal for incorporating dose-toxicity data learnt from animals into human trials. We have presented our proposal adopting a two-parameter logistic regression model to describe dose-toxicity relationship. However, more sophisticated models such as physiologically based pharmacokinetic model[43] may be considered. For the species-appropriate translation parameter introduced in our model, we assume that allometric scaling principles adjusting for body surface area[44,45] adequately describe physiological differences between animals and humans. Additional work would be needed to verify the appropriateness of this approach or refine it, since it may be inappropriate in some circumstances, for example, when the compound is a monoclonal antibody[46] or a biological agent[47]. The approach we proposed here has also some limitations. First and foremost, it relies on the assumption that we can adequately extrapolate from animals to humans. If this assumption is questionable, then the model may fail to correctly translate the animal dose-toxicity information onto the human scale. Furthermore, we use a statistical model (logistic regression) to describe the dose-toxicity relationship, which means that no mechanistic insights into the cause of the toxicities will be gained. Finally, we did not address more subtle differences that often exist between

animal and human studies, for example, mode of administration and handling of drop-outs.

In this paper we specifically focus on the transition step from preclinical to clinical studies in early drug development, but the methodology proposed in Section 2 can be applied more broadly: it can be used to augment a clinical trial with historical data that have been recorded on a different measurement scale. Further research will extend the proposed model to accommodate heterogeneity amongst humans. Potential applications include the case that phase I dose-escalation bridging studies to be carried out in different geographic regions. Alternatively, there may be differences between age groups, for example, between children and adults, or adults and geriatrics.

Supplemental material

Supplementary materials may be found in the online version of this article at the publishers website. The authors have also provided additional supporting information of this article. Specifically, (a) OpenBUGS code to implement the proposed model, together with R functions needed to reproduce the results reported in Sections 3 and 4; (b) R functions used to derive a log-normal prior for the species-appropriate translation factor are provided.

Acknowledgements

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 633567. Dr Hampsons contribution to this manuscript was supported by the UK Medical Research Council (grant MR/M013510/1). The authors would like to thank Dr Beat Neuenschwander for helpful methodological discussions and Dr Michéle Bouisset-Leonard for her insightful comments on practical translational science.

References

1. Viele K, Berry S, Neuenschwander B et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics* 2014; 13(1): 41–54.
2. Neuenschwander B, Roychoudhury S and Schmidli H. On the use of co-data in clinical trials. *Statistics in Biopharmaceutical Research* 2016; 8(3): 345–354.
3. van Rosmalen J, Dejardin D, van Norden Y et al. Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research* 2017; 0(0): 1–16.

4. Eichler H, Pétavy F, Pignatti F et al. Access to patient-level trial data – a boon to drug developers. *New England Journal of Medicine* 2013; 369(17): 1577–1579.
5. Eichler H, Bloechl-Daum B, Bauer P et al. "threshold-crossing": a useful way to establish the counterfactual in clinical trials? *Clinical Pharmacology & Therapeutics* 2016; 100(6): 699–712.
6. Hobbs B, Carlin B and Sargent D. Adaptive adjustment of the randomization ratio using historical control data. *Clinical Trials* 2013; 10(3): 430–440.
7. French J, Wang S, Warnock B et al. Historical control monotherapy design in the treatment of epilepsy. *Epilepsia* 2010; 51(10): 1936–1943.
8. French J, Temkin N, Shneker B et al. Lamotrigine xr conversion to monotherapy: first study using a historical control group. *Neurotherapeutics* 2012; 9(1): 176–184.
9. Wadsworth I, Hampson L and Jaki T. Extrapolation of efficacy and other data to support the development of new medicines for children: a systematic review of methods. *Statistical Methods in Medical Research* 2018; 27(2): 398–413.
10. Dane A and Wetherington J. Statistical considerations associated with a comprehensive regulatory framework to address the unmet need for new antibacterial therapies. *Pharmaceutical Statistics* 2014; 13(4): 222–228.
11. Takeda K and Morita S. Bayesian dose-finding phase I trial design incorporating historical data from a preceding trial. *Pharmaceutical Statistics* 2018; 0(0): 1–11.
12. Cunanan K and Koopmeiners J. Hierarchical models for sharing information across populations in phase I dose-escalation studies. *Statistical Methods in Medical Research* 2017; 0(0): 1–13.
13. Ibrahim J and Chen M. Power prior distributions for regression models. *Statistical Science* 2000; 15(1): 46–60.
14. Duan Y, Smith E and Ye K. Using power priors to improve the binomial test of water quality. *Journal of Agricultural, Biological, and Environmental Statistics* 2006; 11(2): 151.
15. Hobbs B, Carlin B, Mandrekar S et al. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* 2011; 67(3): 1047–1056.
16. Hobbs B, Sargent D and Carlin B. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis* 2012; 7(3): 639–674.
17. Neuenschwander B, Capkun-Niggli G, Branson M et al. Summarizing historical information on controls in clinical trials. *Clinical Trials* 2010; 7(1): 5–18.
18. Schmidli H, Gsteiger S, Roychoudhury S et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 2014; 70(4): 1023–1032.
19. O'Quigley J, Pepe M and Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics* 1990; 46(1): 33–48.
20. Paoletti X and Kramar A. A comparison of model choices for the continual reassessment method in phase I cancer trials. *Statistics in Medicine* 2009; 28(24): 3012–3028.

21. Babb J, Rogatko A and Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine* 1998; 17(10): 1103–1120.
22. Whitehead J. Using Bayesian decision theory in dose-escalation studies. In Chevret S (ed.) *Statistical Methods for Dose-Finding Experiments*, chapter 7. Statistics in Practice, Wiley-Blackwell, 2006.
23. Storer B. Design and analysis of phase I clinical trials. *Biometrics* 1989; 45(3): 925–937.
24. Jaki T, Clive S and Weir C. Principles of dose finding studies in cancer: a comparison of trial designs. *Cancer Chemotherapy and Pharmacology* 2013; 71(5): 1107–1114.
25. Whitehead J and Williamson D. Bayesian decision procedures based on logistic regression models for dose-finding studies. *Journal of Biopharmaceutical Statistics* 1998; 8(3): 445–467.
26. Neuenschwander B, Branson M and Gsponer T. Critical aspects of the Bayesian approach to phase I cancer trials. *Statistics in Medicine* 2008; 27(13): 2420–2439.
27. FDA. *Estimating the Maximum Safe Starting Dose in Initial Clinical Trials for Therapeutics in Adult Healthy Volunteers*. US Food and Drug Administration: Rockville, MD, 2005.
28. Reigner B and Blesch K. Estimating the starting dose for entry into humans: principles and practice. *European Journal of Clinical Pharmacology* 2001; 57: 835–845.
29. Neuenschwander B, Matano A, Tang Z et al. A Bayesian industry approach to phase I combination trials in oncology. In Zhao W and Yang H (eds.) *Statistical Methods in Drug Combination Studies*, chapter 6. Chapman & Hall/CRC Biostatistics Series, CRC Press, 2014.
30. Kamrin M. *Toxicology – A Primer on Toxicology Principles and Applications.* CRC Press, 1988.
31. West G and Brown J. The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. *Journal of Experimental Biology* 2005; 208(9): 1575–1592. DOI:10.1242/jeb.01589.
32. Sharma V and McNeill J. To scale or not to scale: the principles of dose extrapolation. *British Journal of Pharmacology* 2009; 157(6): 907–921.
33. Baker S, Verweij J, Rowinsky E et al. Role of body surface area in dosing of investigational anticancer agents in adults, 1991–2001. *JNCI: Journal of the National Cancer Institute* 2002; 94(24): 1883–1888.
34. Lunn D, Spiegelhalter D, Thomas A et al. The bugs project: Evolution, critique and future directions. *Statistics in Medicine* 2009; 28(25): 3049–3067.
35. Roman D, VerHoeve J, Schadt H et al. Ocular toxicity of auy922 in pigmented and albino rats. *Toxicology and Applied Pharmacology* 2016; 309(Supplement C): 55–62.
36. Sessa C, Shapiro G, Bhalla K et al. First-in-human phase I dose-escalation study of the hsp90 inhibitor auy922 in patients with advanced solid tumors. *Clinical Cancer Research* 2013; 19(13): 3671–3680.
37. Gelman A, Jakulin A, Pittau M et al. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2008; 2(4): 1360–1383.

38. Morita S, Thall P and Müller P. Determining the effective sample size of a parametric prior. *Biometrics* 2008; 64(2): 595–602.
39. Zhou Y and Whitehead J. Practical implementation of Bayesian dose-escalation procedures. *Drug Information Journal* 2003; 37(1): 45–59.
40. O'Quigley J, Paoletti X and Maccario J. Non-parametric optimal design in dose finding studies. *Biostatistics* 2002; 3(1): 51–56.
41. Dresser R. First-in-human trial participants: Not a vulnerable population, but vulnerable nonetheless. *The Journal of Law, Medicine & Ethics* 2009; 37(1): 38–50.
42. Cook N, Hansen A, Siu L et al. Early phase clinical trials to identify optimal dosing and safety. *Molecular Oncology* 2015; 9(5): 997–1007.
43. Gueorguieva I, Aarons L and Rowland M. Diazepam pharamacokinetics from preclinical to phase I using a Bayesian population physiologically based pharmacokinetic model with informative prior distributions in winbugs. *Journal of Pharmacokinetics and Pharmacodynamics* 2006; 33(5): 571–594.
44. Kouno T, Katsumata N, Mukai H et al. Standardization of the body surface area (bsa) formula to calculate the dose of anticancer agents in Japan. *Japanese Journal of Clinical Oncology* 2003; 33(6): 309–313.
45. Gerina-Berzina A, Vikmanis U, Teibe U et al. Anthropometric measurements of the body composition of cancer patients determine the precise role of the body surface area and the calculation of the dose of chemotherapy. *Papers on Anthropology* 2012; 21(0).
46. Department of Health. *Expert Scientific Group on Phase One Clinical Trials: Final Report*. London: HMSO, 2006. URL http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_063117.
47. Tang L, Persky A, Hochhaus G et al. Pharmacokinetic aspects of biotechnology products. *Journal of Pharmaceutical Sciences* 2004; 93(9): 2184–2204.

Appendix

A Specifying a log-normal prior for the translation factor $\delta_{\mathcal{A}_i}$

One common method for extrapolating doses across species is allometric scaling performed on the basis of body surface area (BSA). FDA[27] proposed calculating a human-equivalent dose (HED) by multiplying the animal dose by a factor reflecting the relationship between metabolic rate and mass in mammals:

$$\mathrm{HED(mg/kg)} = \mathrm{Animaldose(mg/kg)} \times \frac{(\mathrm{BW/BSA})_{\mathrm{Animal}}}{(\mathrm{BW/BSA})_{\mathrm{Human}}}, \tag{7}$$

where BW denotes the body weight (kg) and BSA is measured in square metres.

In the notation of this paper, $\delta_{\mathcal{A}_i} = ((\mathrm{BW}/\mathrm{BSA})_{\mathrm{Animal}}/(\mathrm{BW}/\mathrm{BSA})_{\mathrm{Human}})$ is the interspecies translation factor. As noted in Section 2, we fit models treating each δ_{S_k} as a random variable rather than a fixed constant to formally account for uncertainty about translation factors. An independent log-normal prior is placed on each δ_{S_k} consistent with the translation factor in (7). Body weight is commonly modelled by a log-normal distribution, whilst for present purposes, we assume the body surface area has negligible variation in animals and humans. As both the numerator and denominator of (7) are log-normally distributed, the translation factor can be described using a log-normal distribution.

Given the species-specific body weight and body surface area information available from the FDA draft guideline, displayed at the left of Table 1, we derive log-normal priors, based on an optimiser, so that medians and 95% CIs are in good agreement with the reference and working range of body weight. This is an optimisation problem in the sense that we aim to minimise the distance between the summaries (reference and working range) and the key percentiles (2.5th, 50th and 97.5th percentiles) of the log-normal prior. Specifically,

- For each animal species, BW/BSA can be summarised as $Q = \{q_L, q_M, q_U\}$, in which q_M corresponds to the reference value and $[q_L, q_U]$ as the limits of the working range
- The reference value is taken as median of the log-normal prior
- The log-normal variance is determined such that the absolute distance between the implied 2.5th and 97.5th percentiles and q_L and q_U is minimised, respectively
- Likewise, derive the log-normal prior for BW/BSA in humans
- Depending on the unit of human dose, either mg/kg or mg/m^2, the log-normal prior for $\delta_{\mathcal{A}_i}$ is therefore obtained.

R code for the derivation is available at the publisher's website.

B Priors for other parameters

Weakly informative priors for the robust component and the population means $\boldsymbol{m}$:

- Prior for $\theta_{1i^\star}$: $m_{01} \sim N(\log\left(\frac{0.25}{1-0.25}\right), 2^2)$. This suggests that prior median for the probability of toxicity at $d_{\text{Ref}} = 28$ mg/m^2 is 0.25 and the 95% credible interval is (0.007, 0.944).
- Prior for $\theta_{2i^\star}$: $m_{02} \sim N(0, 1^2)$. This prior for the slope parameter is weakly informative as it allows for flat to very steep curves. Under this specification, when doubling the dose, the odds of a DLT is multiplied by $2^{\exp(0)} = 2$ (prior median), and the 95% credible interval for this multiplier is (1.1, 137.1).
- Priors for m_1 and m_2: $m_1 \sim N(\log\left(\frac{0.25}{1-0.25}\right), 1.98^2)$, and $m_2 \sim N(0, 0.99^2)$. These priors are similar to the ones for the robust component and therefore are also weakly informative.

Half-normal distributions are chosen for elements of the covariance matrix Ψ and Σ as follows.

- Priors for τ_1 and τ_2 that control borrowing within same species: $\tau_1 \sim HN(0.5)$, of which the key summaries, say, median and 95% credible interval, are 0.337 and (0.016, 1.121), respectively. This allows for substantial between-study heterogeneity for the intercept parameter, θ_{1i}. $\tau_2 \sim HN(0.25)$, of which the key summaries, say, median and 95% credible interval, are 0.169 and (0.008, 0.560), respectively. This allows for moderate between-study heterogeneity for the slope parameter, θ_{2i}.
- Priors for σ_1 and σ_2 that control borrowing across different animal species: $\sigma_1 \sim HN(15)$, of which the median and 95% credible interval are 10.117 and (0.470, 33.621), respectively; $\sigma_2 \sim HN(5)$, of which the median and 95% credible interval are 3.372 and (0.157, 11.207), respectively. These are diffused priors used in the paper for the special case $K = 1$.
- Priors for the correlation coefficients: $\rho \sim U(-1, 1)$ and $\kappa \sim U(-1, 1)$.

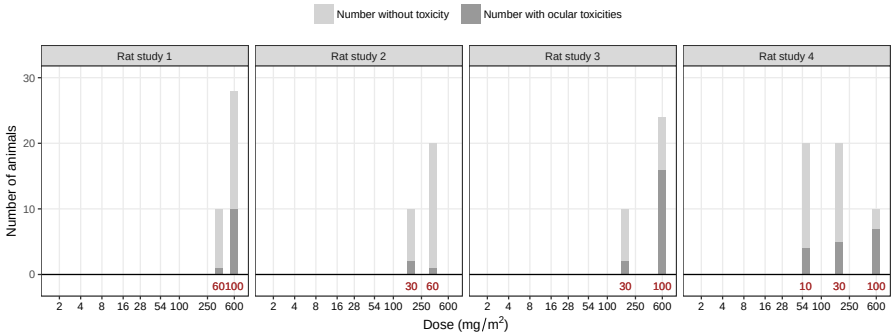

Figure 1. Preclinical data from four studies in rats. The height of the bar represents the number of rats studied, and the height of the dark grey segment counts the number experiencing an ocular toxicity. Doses listed in brown are the doses (mg/kg) administered to rats. Doses listed in black are the human-equivalent doses (mg/m^2). Projections are made by scaling animal doses using the prior median of δ_{Rat}.

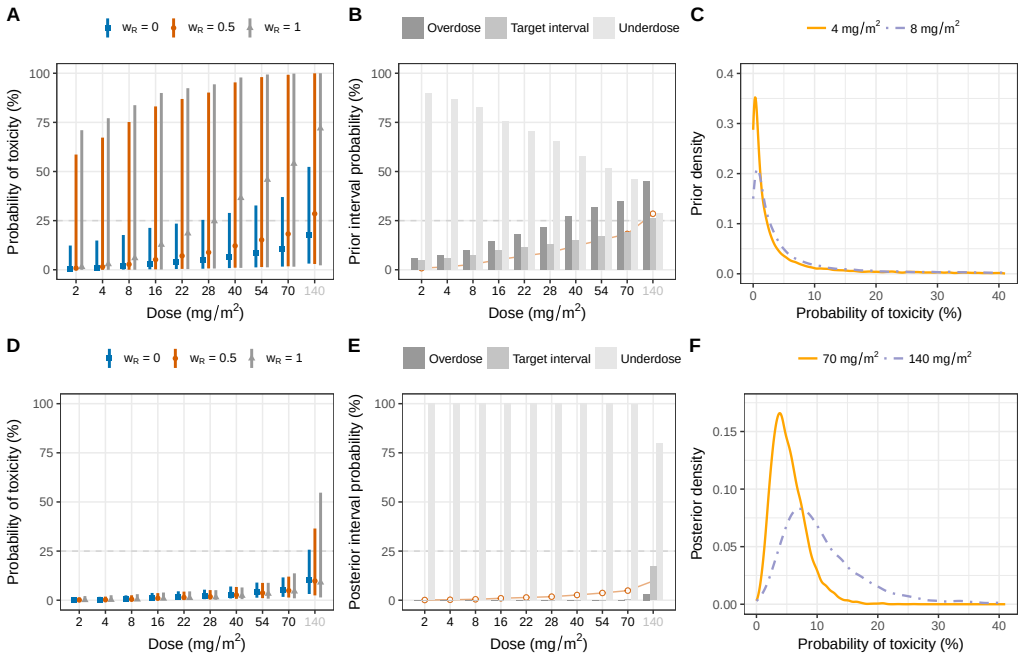

Figure 2. Results of the Bayesian meta-analysis, corresponding to the synthesis of ocular toxicity data in rats without and with the human data, respectively. Panels A and D show median and 95% CI of the marginal distributions for the probability of ocular toxicity. Panels B and E describe the marginal distributions of $w_R = 0.5$ using interval probabilities. The background red curve shows the median probability of toxicity of each human dose. Panels C and and F display the entire marginal distributions for the risk of ocular toxicity on doses of particular interest.

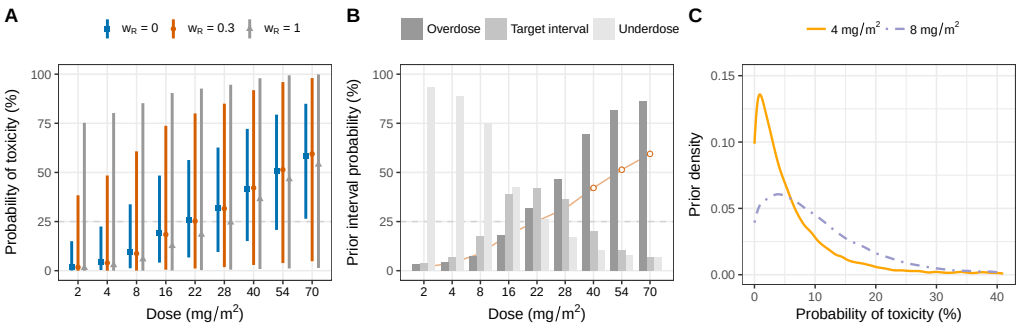

Figure 3. Summaries about the Bayesian analyses of the binary DLT data in dogs. Panel A shows median and 95% CI of the marginal prior predictive distribution for the probability of toxicity in the future human phase I trial, for a range of doses to be assessed. Prior predictive distributions are derived from a Bayesian meta-analysis of the dog data alone, setting w_R = 0, 0.3 or 1. Panel B gives an overview on the toxicity interval probabilities predicted based on a robust meta-analysis of dog data, setting w_R = 0.3. The background red curve shows the prior median probability of toxicity per human dose. Panel C presents prior densities for the risks of toxicity at potential starting doses.

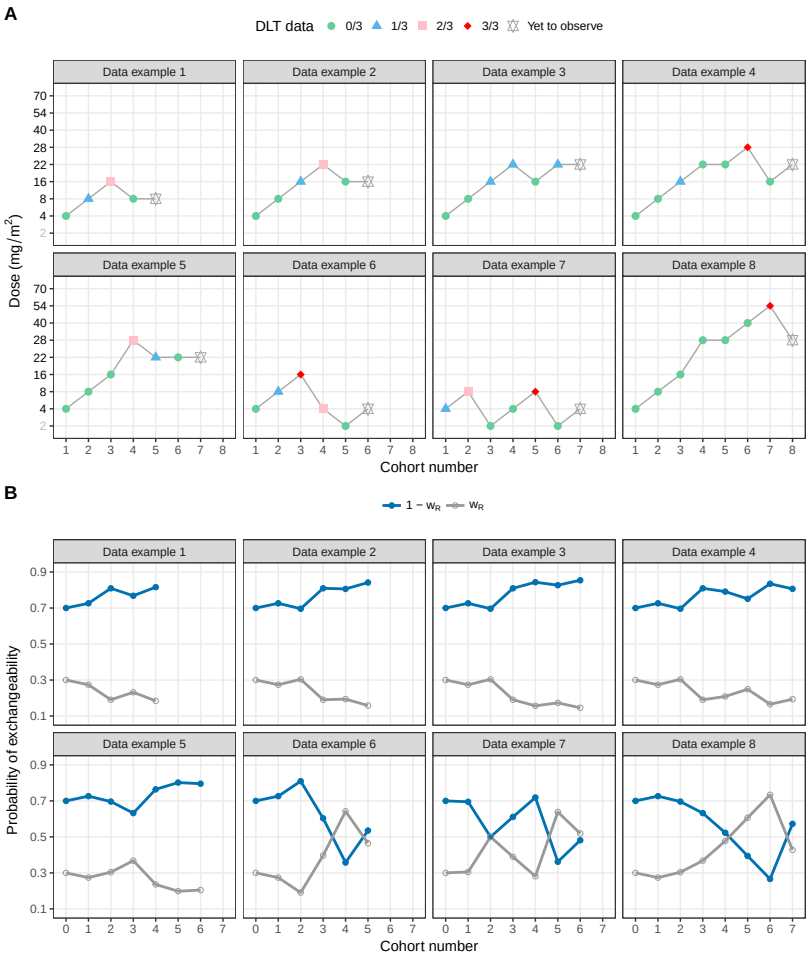

Figure 4. Trajectory of dose recommendations (Panel A) and posterior probabilities of exchangeability (Panel B) during the course of each hypothetical phase I trial in data examples 1 to 8.

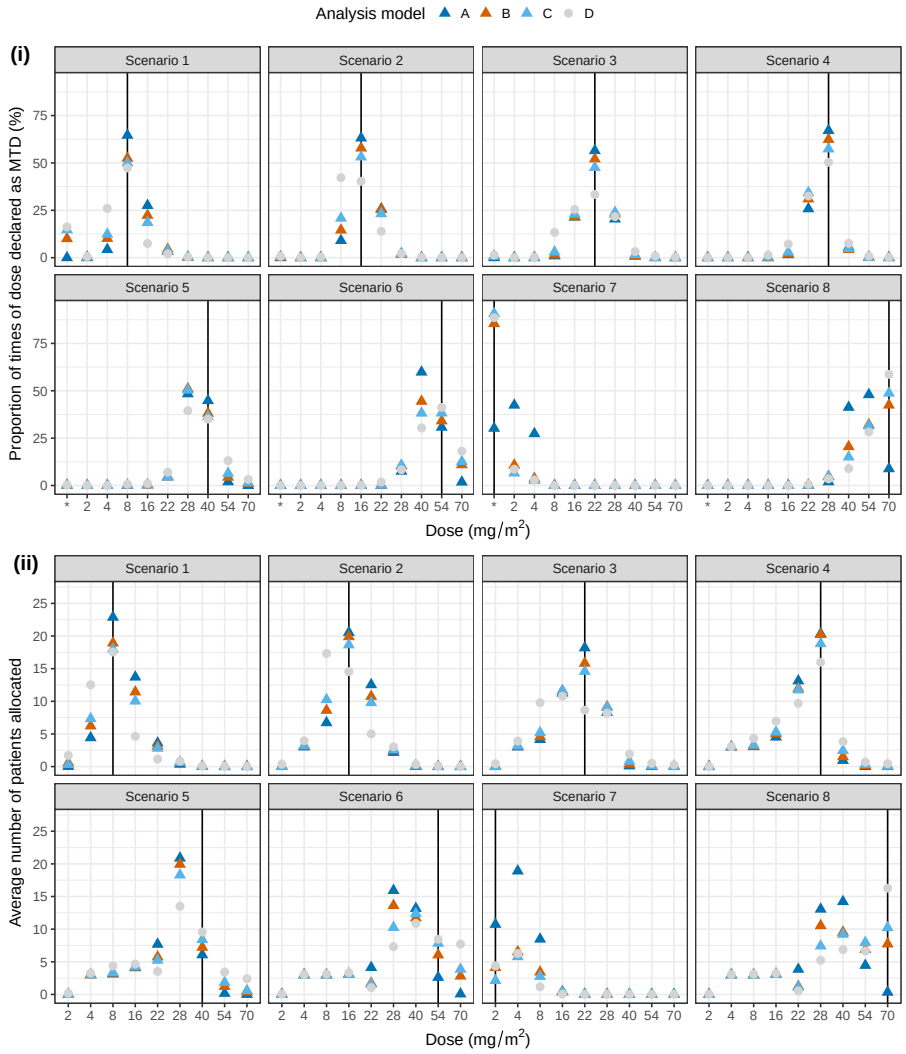

Figure 5. Operating characteristics of BLRM-guided dose-escalation procedures basing inferences on Models A-D, defining δ_{Dog} as a random variable. The vertical black line indicates the true MTD in humans in each simulation scenario.

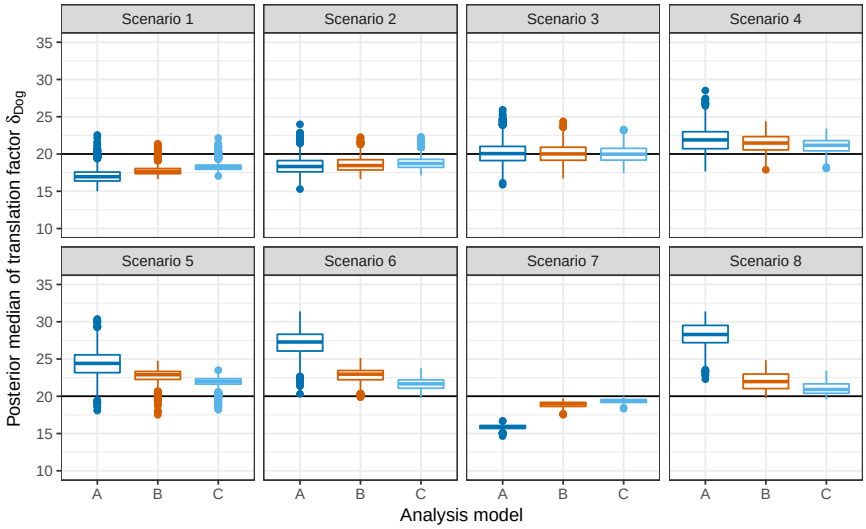

Figure 6. Boxplots of poesterior medians of the translation parameter δ_{Dog} under each meta-analytic model over all *completed* trials. The horizontal black line represents the prior median of δ_{Dog}.